

Model selection and parameter inference in phylogenetics using Nested Sampling

Patricio Maturana R.^{1*} Brendon J. Brewer¹ Steffen Klaere¹²

¹Department of Statistics, University of Auckland, New Zealand

²School of Biological Sciences, University of Auckland, New Zealand

March 17, 2017

Abstract

Bayesian inference methods rely on numerical algorithms for both model selection and parameter inference. In general, these algorithms require a high computational effort to yield reliable inferences. One of the major challenges in phylogenetics regards the estimation of the marginal likelihood. This quantity is commonly used for comparing different evolutionary models, but its calculation, even for simple models, incurs high computational cost. Another interesting challenge regards the estimation of the posterior distribution. Often, long Markov chains are required to get sufficient samples to carry out parameter inference, especially for tree distributions. In general, these problems are addressed separately by using different procedures. Nested sampling (NS) is a Bayesian algorithm which provides the means to estimate marginal likelihoods and to sample from the posterior distribution at no extra cost. In this paper, we introduce NS to phylogenetics. Its performance is analysed under different scenarios and compared to established methods. We conclude that NS is a very competitive and attractive algorithm for phylogenetic inference.

Key Words: model selection, parameter inference, nested sampling, marginal likelihood

1 Introduction

Bayesian methods provide a comprehensive framework in which to explore parameter space, uncertainty and model-to-data fitness. The concept was introduced to phylogenetics in the 1990s (Rannala and Yang, 1996; Yang and Rannala, 1997; Larget and Simon, 1999), and has gained popularity because of its flexibility when dealing with complex models and large data sets, in contrast with maximum likelihood estimation. The increase in computational power led to the rise of Bayesian methods, as Markov Chain Monte Carlo (MCMC) approaches became feasible. Its popularity was further increased by state-of-the-art implementations of the models in programs like MrBayes (Huelsenbeck and Ronquist, 2001), BEAST (Drummond and Rambaut, 2007), and PhyloBayes (Lartillot et al., 2009).

As in other fields, model selection plays an integral part in phylogenetic inference. A wide variety of different criteria are available for this task, with some of the most popular being the Likelihood Ratio Test (LRT), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Bayes Factors (BF), which are ratios of marginal likelihoods. The latter is of particular interest as it provides many advantages over the other methods: i) It is a direct consequence of probability theory used as a theory of reasoning; ii) it allows

*Corresponding author. Address for correspondence: Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. Email address: p.russel@auckland.ac.nz (Patricio Maturana R.)

comparison of nested and non-nested models; iii) it is not based on a point estimate in parameter space since it averages over parameter space; and iv) it embodies Occam’s razor by involving the prior distribution in the model selection process. Marginal likelihoods penalize the inclusion of a new parameter when its value is unknown and some of the possible values do not fit the data well. However, the marginal likelihood is a difficult integral that depends on the complexity of the model. Finding ways to efficiently estimate this integral is one of the major challenges of the field.

A simple Monte Carlo method for estimating the marginal likelihood is the harmonic mean (Newton and Raftery, 1994). However, it is well-known to overestimate the real value, and in many situations its variance is infinite. Among the most accurate methods currently used in phylogenetics are thermodynamic integration (Lartillot and Philippe, 2006) and steppingstone sampling (Xie et al., 2011) which are much more accurate but have a high computational cost. The latter has gained popularity in recent years due to its implementation in different phylogenetic software packages. However, these methods have a relatively large number of tuning parameters that need to be set prior to analysis, and there is no rigorous method of determining the values appropriate for the accurate estimation of the marginal likelihood. Also, these methods have problems dealing with partly convex likelihood functions (Skilling, 2006).

To efficiently deal with the above issues a generalised version of stepping stone sampling (GSS) has been proposed (Fan et al., 2011). This generalization also allows us to regard the phylogeny as an unknown parameter (Holder et al., 2014) incorporating the uncertainty in the tree topology in model selection.

A more general technique for the estimation of the marginal likelihood is nested sampling (NS; Skilling, 2006). This method requires less tuning and can deal with partly convex likelihood functions. Its main feature is the reduction of the multidimensional integral over parameter space to a one-dimensional integral of the likelihood as a function of the enclosed prior probability. This technique, and several variants (e.g. Feroz et al., 2009; Brewer et al., 2011; Handley et al., 2015) have been successfully applied to fields like astronomy (Mukherjee et al., 2006; Brewer and Donovan, 2015) and systems biology (Aitken and Akman, 2013; Pullen and Morris, 2014) and have shown great promise in parameter inference and model selection.

In this paper, we assess the merits of nested sampling to phylogenetic inference, and compare it to established methods. Firstly, we assess its marginal likelihood estimate in a manageable scenario. Secondly, we use it to discriminate among 6 evolutionary models. Then, we test the sensitivity of the estimate to prior specifications for the selected substitution model and we also evaluate its estimated tree posterior distribution. Finally, we carry out tree parameter inference in a bigger data set with a challenging posterior distribution.

2 Bayesian inference

Let θ be the vector of parameters, \mathbf{X} the data, and M the model (assumed throughout). Bayes’ theorem is given by

$$f(\theta|\mathbf{X}, M) = \frac{L(\mathbf{X}|\theta, M)\pi(\theta|M)}{f(\mathbf{X}|M)}. \quad (1)$$

The prior distribution $\pi(\theta|M)$ represents our previous knowledge of the parameters which is updated after taking into account the data; the updated knowledge is reflected in the posterior probability distribution $f(\theta|\mathbf{X}, M)$. The likelihood function $L(\mathbf{X}|\theta, M)$ represents the probability of the data given the parameters and the model. The marginal likelihood $f(\mathbf{X}|M)$ is the probability of the data under the model and plays a key role in model selection. Indeed, this quantity is used to select among models. Because of this, it is also called **the evidence** (MacKay, 2002). To understand its role, note that the posterior distribution for the model M_i is given by

$$f(M_i|\mathbf{X}) = \frac{f(\mathbf{X}|M_i)f(M_i)}{f(\mathbf{X})}, \quad i = 0, 1,$$

where $f(\mathbf{X}|M_i)$ is the marginal likelihood as defined in (1), $f(M_i)$ is the prior probability for the model, and $f(\mathbf{X})$ is the probability of the data. The marginal likelihood will also be denoted by \mathcal{Z} henceforth. The

Bayesian comparison of two models M_0 and M_1 can be carried out by comparing their posterior probabilities. This comparison is often through the ratio of their probabilities, which represents the plausibility of one model over another and is defined as follows:

$$\frac{f(M_0|\mathbf{X})}{f(M_1|\mathbf{X})} = \frac{f(\mathbf{X}|M_0) f(M_0)}{f(\mathbf{X}|M_1) f(M_1)},$$

posterior odds = Bayes factor \times prior odds.

The ratio of marginal likelihoods, the first one of the right side, is called the Bayes factor (Kass and Raftery, 1995). If we have no preference for any model, i.e. each model is assigned the same prior probability, the priors are canceled and the posterior odds is only given by ratio of the marginal likelihoods. Here lies the importance of the latter.

Although the marginal likelihood is often ignored, it plays a key role in model selection: it is a measure of the goodness of fit. Indeed, it is the probability of the data given the model, i.e. it is by definition a measure of model fit. This quantity is a multidimensional integral of the prior distribution times the likelihood function over the parameter space. The marginal likelihood acts as the normalization constant in the posterior distribution making it a probability density function. MCMC methods used for parameter estimation within a model use only ratios of posterior densities, and are therefore unable to measure its normalization in general.

Unlike maximum likelihood, which represents the model fit at a single point, this quantity stands for an average of how well the model fits the data. By being an average of the likelihood function with respect to the prior, the model with the greatest evidence might be different from the model with the highest likelihood because the prior could downweight some regions of parameter space. Also, the marginal likelihood is sensitive to the size of the region over which the likelihood is high. As a result, both methods could favour different models. Despite its important role in model selection, the marginal likelihood is usually analytically intractable and has to be approximated by numerical methods.

2.1 Estimation of marginal likelihoods

Typically, phylogenetic models involve a high level of complexity, making it difficult to calculate the marginal likelihood. Suchard et al. (2001) proposed the Savage-Dickey ratio to estimate Bayes factors for nested models (Verdinelli and Wasserman, 1995). Huelsenbeck et al. (2004) used reversible jump Markov chain Monte Carlo including all possible time-reversible models. Nevertheless, these methods are restricted to a particular group of models. Other alternatives have been proposed to allow a more general comparison of models. Among them, the harmonic mean (HM) is the most popular to estimate the marginal likelihood (Newton and Raftery, 1994), an importance-sampling approach. Its popularity is due to its simplicity, it only requires samples from the posterior distribution. However, the HM estimator often has infinite variance and overestimates the true value of the marginal likelihood (Lartillot and Philippe, 2006; Xie et al., 2011).

Far more accurate than the HM method is thermodynamic integration (TI), proposed by Lartillot and Philippe (2006). The method requires several Markov chains from transition distributions which form a path between the prior and the posterior distribution. These transition functions are defined by the ‘power posterior’

$$p_\beta = \frac{L(\mathbf{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)}{\mathcal{Z}_\beta}, \quad \text{for } 0 \leq \beta \leq 1,$$

where \mathcal{Z}_β is the normalizing constant of the unnormalized power posterior density $L(\mathbf{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)$. Similarly, a path between the posterior of two models could be defined to estimate the Bayes factor directly. Note that for $\beta = 0$ the power posterior is equivalent to the prior distribution and for $\beta = 1$ is equivalent to the posterior distribution. In the latter case, $\mathcal{Z}_1 = \mathcal{Z}$ is the marginal likelihood. TI relies on the identity

$$\log \mathcal{Z} = \int_0^1 \mathbb{E}_{p_\beta} [\log L(\mathbf{X}|\boldsymbol{\theta}, M)] d\beta,$$

where the expected value is with respect to the power posterior distribution. TI uses a series of β values which define the transition distributions. For each value, a Markov chain is required to estimate the expected values and consequently the integral over β . Clearly, the increase in accuracy comes at the cost of increased complexity.

Another importance sampling approach is stepping-stone sampling (SS) proposed by [Xie et al. \(2011\)](#). SS relies on transition distributions like TI in order to define an equivalence between the marginal likelihood and the telescope product of ratios of normalizing constants given by

$$\mathcal{Z} = \prod_{k=1}^K \frac{\mathcal{Z}_{\beta_k}}{\mathcal{Z}_{\beta_{k-1}}},$$

where $\beta_0 = 0 < \beta_1 < \dots < \beta_{K-1} < \beta_K = 1$. Each ratio $\mathcal{Z}_{\beta_k}/\mathcal{Z}_{\beta_{k-1}}$ is estimated by importance sampling. The performance of this method is similar to TI. However, SS requires slightly less computational effort: it does not need posterior samples and in general requires a smaller number of transition distributions. SS also allows us to estimate the Bayes factor directly defining a path between the posterior of both models ([Baele et al., 2013](#)). The extended version of SS, generalized steppingstone sampling (GSS; [Fan et al., 2011](#)), uses a reference distribution to shorten the distance between the prior and posterior distribution. This strategy can potentially lead to a more efficient estimation process. Like in its predecessors, the geometric path is often used to connect these densities which is defined by

$$p_\beta = \frac{(L(\mathbf{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M))^\beta \pi_0(\boldsymbol{\theta}|M)^{1-\beta}}{\mathcal{Z}_\beta}, \quad \text{for } 0 \leq \beta \leq 1,$$

where π_0 is the reference distribution. When $\beta = 0$ and $\beta = 1$, the power posterior is equivalent to the reference and posterior distribution, respectively. GSS is more accurate than the original version when the reference distribution approximates the posterior distribution reasonably well. This method has also been extended, allowing the tree topology to be variable ([Holder et al., 2014](#)). This allows the user to accommodate phylogenetic uncertainty in model selection.

Methods to estimate the marginal likelihood should be sensitive to the prior choice. Non-informative priors should increase the contribution of low-likelihood regions of parameter space in the estimated marginal likelihood. Consequently, the prior choice should affect the estimated evidence. However, the harmonic mean is dominated by the posterior density and thus ignores low-likelihood regions. As a result, this method is insensitive to priors, which explains its poor performance. This was shown by [Xie et al. \(2011\)](#) who also showed that thermodynamic integration and steppingstone sampling are sensitive to the assumed prior distribution.

Although TI and SS yield accurate estimates of the marginal likelihood, they require several specifications depending on the problem. Firstly, an annealing schedule (a number of β -values) is required. A common practice is to try with different numbers until the estimate is stable. This procedure is described in [Drummond and Bouckaert \(2015\)](#) as follows: “*run the path sampling analysis with a low number of steps (say 10) first, then increase the number of steps (with say increments of 10, or doubling the number of steps) and see whether the marginal likelihood estimates remain unchanged*”. This could be impractical in some situations, for instance, when flat priors are used, which would increase the number of steps. Secondly, the path described by the β -values has to be defined. [Lartillot and Philippe \(2006\)](#) proposed to spread the β -values regularly spaced between 0 and 1. But since often most of the variability of the expected values is concentrated for β near 0, some authors have proposed to concentrate the computational effort in that place. For example [Lepage et al. \(2007\)](#) used a sigmoidal function to estimate the Bayes factor using TI; [Friel and Pettitt \(2008\)](#) proposed $\beta_k = x_k^4$ in TI, where x -values are equally spaced between 0 and 1; and [Xie et al. \(2011\)](#) advocated spreading the values according to evenly spaced quantiles of a $\text{Beta}(\alpha, 1)$, with $\alpha = 0.3$. Finally, these methods require a number of samples from the power posterior for each β -value. Thus, the main problem is that optimal specifications vary from case to case. The popularity of SS is due to its implementation in popular software such as MrBayes ([Huelsenbeck and Ronquist, 2001](#)) or BEAST ([Drummond et al., 2012](#)). However the mentioned specifications have to be defined by the user, or use some predetermined parameters that might be unsuitable.

In this context, GSS requires potentially less tuning parameters for an appropriate reference distribution. Firstly, it requires an annealing/melting scheme (a number of β -values). The estimation can start from either the prior or posterior distribution. However, the β -values do not require to follow any particular distribution to control effectively the uncertainty of the estimate as in TI or SS, because of the similarity of the reference and posterior distributions. Thus, the values can be equally spaced between 0 and 1. Also, GSS does not need as many transitional distributions as its original version and it is more robust to prior specifications, i.e. the prior does not have a huge effect on the method performance. Finally, the method requires a number of samples from each transitional distribution.

TI and SS have usually been presented as methods of general applicability (Xie et al., 2011; Baele et al., 2013; Arima and Tardella, 2014; Baele and Lemey, 2014). However, these methods only work when the shape of the likelihood function is concave. Partly convex likelihood functions might need impractical computational effort or make them fail outright. The transition distributions are unable to mix between different phases of the likelihood function, resulting in a poor estimate. A more general method is nested sampling (Skilling, 2006), an algorithm that measures the relationship between likelihood values and the prior distribution, and uses this to compute the marginal likelihood. This characteristic allows it to cope with partly convex likelihood functions. More importantly, unlike TI, SS and GSS, NS requires less problem-specific tuning.

3 Nested Sampling

The marginal likelihood or evidence is given by

$$\mathcal{Z} = \int_{\Theta} \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (2)$$

where $\boldsymbol{\theta} \in \Theta$ is the parameter vector, $L(\boldsymbol{\theta})$ is the likelihood function and $\pi(\boldsymbol{\theta})$ is the prior distribution. All the conditionals have been omitted, namely, $L(\boldsymbol{\theta})$ is used as the likelihood function instead of $L(\boldsymbol{\theta}|\mathbf{X}, M)$ and $\pi(\boldsymbol{\theta})$ is used as the prior instead of $\pi(\boldsymbol{\theta}|M)$.

This definition applies for a continuous parameter space Θ . This is the case when the phylogeny is fixed. However, when the tree topology is unknown, the parameter space is additionally composed by a discrete part. In this case, the marginal likelihood incorporates the sum over the tree parameter space and is known as *total marginal likelihood*. Strictly speaking, its definition is given by

$$\mathcal{Z} = \sum_{\tau \in \mathcal{T}} \int_{\mathcal{V}_{\tau}} \int_{\Theta} L(\mathbf{X}|\boldsymbol{\theta}, \nu_{\tau}, \tau, M) \pi(\boldsymbol{\theta}, \nu_{\tau}, \tau|M) d\boldsymbol{\theta} d\nu_{\tau},$$

where $\boldsymbol{\theta} \in \Theta$ is the parameter vector composed by elements such as frequencies, gamma parameter and rates parameters, $\nu_{\tau} \in \mathcal{V}_{\tau}$ is the set of branch lengths of $\tau \in \mathcal{T}$ which is the tree topology, \mathbf{X} is the molecular data and M is the substitution model. For simplicity, NS will be described for definition (2), but its generalization to the case of variable tree topology is analogous.

To understand the key idea of NS, consider that for any positive random variable Y , its expected value can be written as

$$\mathbb{E}[Y] = \int_0^{\infty} (1 - F(Y)) dY,$$

which depicts the area between the distribution function of Y and 1. Similarly, the likelihood function $L(\boldsymbol{\theta})$ can be seen as a positive random variable where $\boldsymbol{\theta}$ follows the prior distribution $\pi(\boldsymbol{\theta})$ and the evidence as the expected value of the likelihood function. Thus, nested sampling takes advantage of this property to transform the multi-dimensional integral defined in (2) into a one-dimensional integral as follows

$$\mathbb{E}_{\boldsymbol{\theta}}[L(\boldsymbol{\theta})] \equiv \mathbb{E}_{\lambda}[\lambda] = \int_0^{\infty} (1 - F(\lambda)) d\lambda, \quad (3)$$

where $\mathbb{E}_\theta[\cdot]$ and $\mathbb{E}_\lambda[\cdot]$ stand for the expectation with respect to the densities of θ and λ respectively, $\theta \sim \pi(\theta)$, $\lambda = L(\theta)$ and $F(\lambda)$ is the cumulative distribution function of the likelihood defined by

$$F(\lambda) = \int \cdots \int_{L(\theta) < \lambda} \pi(\theta) d\theta.$$

Considering $\xi(\lambda) = 1 - F(\lambda)$, the proportion of prior mass with likelihood greater than λ , and taking its inverse, the evidence given in (3) is redefined as

$$\mathcal{Z} = \int_0^1 L(\xi) d\xi.$$

This is the integral used by nested sampling, and is displayed in Figure 1. In general, this function concentrates its mass near zero because the posterior is located in a small area of the prior. We use the “overloaded” notation, where the same letter L represents the likelihood function over different domains: $L(\theta)$ has the parameter vector θ as argument, and $L(\xi)$ has the prior mass ξ (scalar) as argument. Note that $L(\xi)$ is a monotonically decreasing function which reaches its highest point at $\xi = 0$ and its lowest point at $\xi = 1$ (see Figure 1). $L(0.9) = 0.3$ means that 90% of the draws θ from the prior distribution will have likelihoods greater than 0.3. If a set of points on the $L(\xi)$ curve can be obtained, the integral can be approximated numerically by the basic standard quadrature method

$$\mathcal{Z} \approx \sum_{i=1}^k w_i L_i, \quad (4)$$

where $w_i = \xi_{i-1} - \xi_i$ and $L_i = L(\xi_i)$. For a decreasing sequence of ξ -values and an increasing sequence of L -values the evidence can be estimated. How to generate these sequences is described below.

3.1 Sequence of L -values

Nested sampling maintains a set of N active points $\theta_1, \dots, \theta_N$ (with respective associated likelihood values $L(\theta_1), \dots, L(\theta_N)$) to generate the i th likelihood value required in (4). Initially they are drawn from the prior distribution, $\pi(\theta)$. From this set, the method requires selecting the point θ_l , where $l \in \{1, \dots, N\}$, with the lowest likelihood value. This value contributes to the estimation as a summand in (4). Then, the point θ_l is discarded from the active points and replaced by a new point θ sampled from the prior, but constrained to have a greater likelihood value than the point being replaced, i.e. $L(\theta) > L(\theta_l)$. This procedure shrinks the parameter space according to the likelihood restriction. The process is repeated until a given stopping rule is satisfied (more information on this will follow later). Thus, a sequence of increasing likelihood values (L_1, \dots, L_k) and *discarded points* $(\theta_1, \dots, \theta_k)$ are generated. The discarded points are the ones that contribute to the estimate of the marginal likelihood through their respective likelihoods.

3.2 Sequence of ξ -values

The discarded points generate an increasing sequence of likelihoods, which are known precisely. An important insight of Skilling (2006) is that the corresponding ξ values, while they cannot be measured precisely, can be estimated from the nature of the NS procedure. Nested sampling explores the prior distribution geometrically as follows

$$\xi_0 = 1, \quad \xi_1 = t_1, \quad \xi_2 = t_1 t_2, \quad \dots, \quad \xi_k = \prod_{i=1}^k t_i,$$

where $t_i = \xi_i / \xi_{i-1} \in [0, 1]$, for $i = 1, \dots, k$. This variable follows a Beta($N, 1$) distribution. This is because at the i^{th} iteration, NS takes N x_i points which follows a Uniform($0, \xi_{i-1}$), with $i = 1, \dots, N$. These values are



Figure 1: Association between the cumulative prior mass and the likelihood function. Nested sampling estimates the gray area which is the marginal likelihood. In general, a small area of the prior concentrates high likelihood values making the area be concentrated around $\xi \approx 0$.

cumulative probabilities and consequently have a uniform distribution. Their maximum value is ξ_i which is related to the minimum likelihood value (note that $L(\xi)$ is a non-increasing function). Since the distribution of x_i/ξ_{i-1} is a $\text{Uniform}(0, 1)$, their maximum value ξ_i/ξ_{i-1} follows a $\text{Beta}(N, 1)$ distribution. Skilling (2006) defined two schemes for estimating the ξ -values: *stochastic* and *deterministic*.

- *Stochastic*: the t_i values are generated randomly from the $\text{Beta}(N, 1)$ distribution, for $i = 1, \dots, k$.
- *Deterministic*: the t_i values are fixed by using their expectations as follows:
 - Considering its *arithmetic mean*, $t_i = N/(N + 1)$, approximate ξ -values would be given by

$$\xi_i = \left(\frac{N}{N + 1} \right)^i.$$

- Considering its *geometric mean*, $t_i = e^{-1/N}$, the estimated prior mass would be

$$\xi_i = e^{-i/N}.$$

Thus, a sequence of ξ values can be generated and used in (4). The use of the geometric mean seems more reasonable given that the prior mass exploration is geometric. This scheme is considered for our examples, and is the one recommended by most authors. However, the arithmetic mean allows nested sampling to be connected to rare event simulation (Walter, 2014), and allows for an alternate version of NS with unbiased estimates of \mathcal{Z} .

3.3 Sampling

The highest cost of nested sampling is in sampling from the restricted prior distribution (due to the condition that the likelihood needs to increase). [Skilling \(2006\)](#) suggested to use a Metropolis-Hastings algorithm as usual, to explore the prior with the additional condition of rejecting the proposal points which do not fulfill the likelihood restriction. As a starting value, a point from the sequence of active points can randomly be selected at each iteration of NS, as all of them meet the likelihood condition by definition. Several other efficient methods have also been proposed ([Mukherjee et al., 2006](#); [Feroz et al., 2009](#); [Brewer et al., 2011](#)). We use [Skilling](#)’s method to generate the restricted prior samples in our application.

Unlike the proposal mechanisms used in standard MCMC methods to sample the posterior, a static distribution, in NS such mechanisms have to deal with a variable target distribution over time. In particular, this is a new scenario for tree proposals. Nested sampling compresses the prior at each iteration making it vary at a constant rate. The proposals have to explore a quite wide area at the beginning which becomes constrained over time. Tree proposal mechanisms should be able to adapt to this sampling characteristic. Frequently, a uniform prior distribution is assigned over the tree parameter space which is quite huge even for few taxa. Initially, bold moves would allow a good exploration using less steps than conservative ones. However, the acceptance probability would decrease drastically over time due to the fact that the target distribution gets constrained. On the other hand, conservative moves would require more steps to explore the prior distribution at the beginning, but later on, the acceptance probability would be higher than bold moves. Ideally, the proposal mechanism should take into account this dynamical behaviour of the target distribution over time. Instead of trying to adapt the proposal according to the restricted prior at each iteration, we propose to use a mix between bold and conservative moves generated by *Random Subtree Pruning and Regrafting* (rSPR) and the *Stochastic Nearest Neighbour Interchange* (stNNI) method, respectively.

3.4 Information

The idea of how much we have learned from the data is quantified through the notion of entropy. The measure of information ([Sivia and Skilling, 2006](#); [Knuth and Skilling, 2012](#)) is given by the negative relative entropy

$$H = \int P(\boldsymbol{\theta}) \log \left(\frac{P(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right) d\boldsymbol{\theta},$$

where $P(\boldsymbol{\theta})$ is the posterior distribution. This quantity represents the amount of information in the posterior with respect to the prior, after acquiring data. By definition, it can be seen as the expected value $H = \mathbb{E}_P[\log(P(\boldsymbol{\theta})/\pi(\boldsymbol{\theta}))]$. Its approximation is given by

$$H \approx \sum_i \frac{w_i L_i}{\bar{Z}} \log \left(\frac{L_i}{\bar{Z}} \right)$$

with $w_i = \xi_{i-1} - \xi_i$ ([Sivia and Skilling, 2006](#)). The following property of expected values is useful to understand the use of this concept. If G_Y is the geometric mean of Y , we have that

$$\log G_Y = \mathbb{E}[\log Y] \Leftrightarrow G_Y = e^{\mathbb{E}[\log Y]}. \quad (5)$$

According to this property, e^{-H} is a measure of central tendency or a typical value of $\pi(\boldsymbol{\theta})/P(\boldsymbol{\theta})$. This value can be seen as the bulk of the posterior mass that occupies the prior. This idea helps to define a termination condition for nested sampling which will be described later.

Note that a prior distribution which is consistent with the likelihood function, namely they support the same parameter values, has a lower information than one which likelihood function is in contradiction with the prior, i.e. their mass is concentrated in different places. In other words, if the previous belief changes a lot after acquiring the data, more information has been gained from the data.

The numerical uncertainty of the estimation \mathcal{Z} is approximated in terms of the information and is given by

$$\text{dev}[\log \mathcal{Z}] = \sqrt{\frac{H}{N}}. \quad (6)$$

This estimate does not take into account the error imposed by the integration rule. However, this error is at most $\mathcal{O}(N^{-1})$, and thus negligible in comparison to (6) (Skilling, 2006). The asymptotic variance of the nested sampling approximation grows linearly with the dimension of $\boldsymbol{\theta}$ and its distribution is asymptotically Gaussian (Chopin and Robert, 2010).

3.5 Algorithm

The algorithm iterates between the following steps:

1. Sample N points $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ from the prior $\pi(\boldsymbol{\theta})$;
2. Initialize $\mathcal{Z} = 0$ and $\xi_0 = 1$;
3. Repeat for $i = 1, \dots, k$;
 - i) out of the N live points, take the one with the lowest likelihood which we call $\boldsymbol{\theta}_l$ with corresponding likelihood $L_i = L(\boldsymbol{\theta}_l)$;
 - ii) set $\xi_i = \exp(-i/N)$;
 - iii) set $w_i = \xi_{i-1} - \xi_i$;
 - iv) update $\mathcal{Z} = w_i L_i + \mathcal{Z}$; and
 - v) update the set of active points $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ replacing $\boldsymbol{\theta}_l$ by drawing a new point $\boldsymbol{\theta}$ from the prior distribution restricted to $L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_l)$.

Repeat the routine until a given stopping criterion is satisfied. However, there is not a rigorous criterion that guarantees that we have found most of the bulk of \mathcal{Z} . Nevertheless, some termination conditions have been proposed (Skilling, 2006).

3.5.1 Termination

The loop could continue until the potential maximum new contribution $L_i w_i$ represents a small fraction γ of the accumulated evidence, i.e.

$$\max(L(\boldsymbol{\theta}_1), L(\boldsymbol{\theta}_2), \dots, L(\boldsymbol{\theta}_N)) w_i < \gamma \mathcal{Z}.$$

The algorithm would be stopped when the potential maximum new contribution is not significant.

Another criterion is based on the concept of information defined before. Typically, the likelihood values L start dominating the prior mass w , so the contribution Lw increases at the beginning until the prior mass dominates this quantity. After reaching a maximum, these values start to decrease. The peak of this function is reached in the region of $\xi \approx e^{-H}$, when most of the posterior mass in the prior has been found. Given that $\xi_i \approx e^{-i/N}$, a natural termination condition to estimate the log-evidence would be stopping the loop when i/N significantly exceeds H , i.e., when the posterior mass has been explored completely.

There is no guarantee *in general* that these termination conditions will work perfectly. L might start increasing at a greater rate in the future, overwhelming the points that currently have high weights. In specific cases where the maximum likelihood value is known or can be roughly anticipated, it is possible to be confident that this won't happen.

3.6 Posterior samples

NS yields posterior samples at no extra cost, if we assign appropriate weights to the discarded output points. In each iteration, NS has taken out a point from the active points generating a sequence of discarded points $\theta_1, \theta_2, \dots, \theta_k$. These discarded points have contributed to estimate the marginal likelihood with their respective weights wL which are proportional to the posterior distribution, i.e. prior multiplied by likelihood. Thus, the sequence of discarded points can be sampled according to these weights in order to get a posterior sample. The effective sample size is related to the entropy of the posterior weights as

$$M = \exp \left(- \sum_{i=1}^m p_i \log p_i \right), \quad \text{where} \quad p_i = \frac{w_i L_i}{\mathcal{Z}}.$$

4 Application

We study a statistical model, which was analyzed previously by [Skilling \(2006\)](#), to show a scenario in which many established methods do not work. Further, we analyze 3 phylogenetic data sets to assess nested sampling performance in model selection, parameter inference, and its sensitivity to prior specifications.

We use Metropolis-Hastings to sample from the restricted prior. The prior mass is estimated by using the geometric mean, the deterministic approach. In the phylogenetic examples, we use 50% of rSPR and 50% of stNNI moves as tree proposals. Regarding the prior distributions, we consider the following hierarchical structure for the branch lengths:

$$\begin{aligned} t_i | \mu &\sim \text{Exp}(1/\mu), \quad \text{for } i = 1, \dots, n, \\ \mu &\sim \text{Inverse-Gamma}(\tilde{\alpha}, \tilde{\beta}), \end{aligned}$$

where n is the number of branches for the particular tree topology. [Rannala et al. \(2011\)](#) suggested $\tilde{\alpha} = 3$ and $\tilde{\beta} = 0.2$ in order to prevent an overestimation of the branch lengths. The JC69 model only has these free parameters. For the rate parameters we use

$$\begin{aligned} q_i | \phi &\sim \text{Exp}(\phi), \\ \phi &\sim \text{Exp}(1) \end{aligned}$$

with $i = 1$ for the HKY85 models and $i = 1, 2, 3, 4, 5$ for the GTR models. A Dirichlet(1,1,1,1) is selected for the joint prior of the four nucleotide frequencies in the HKY85 and GTR models and a Gamma(α, β) distribution for the gamma parameter in the HKY85+ Γ_4 , JC69+ Γ_4 , and GTR+ Γ_4 models. The parameters of the latter vary depending on the problem and are defined where applicable. Finally, we use a discrete Uniform for the tree topologies.

All the analyses were done using R ([R Core Team, 2015](#)).

4.1 Statistical example

[Skilling \(2006\)](#) pointed out that thermodynamic integration does not work when the likelihood function is not concave. Such a problem seems like it ought to be easy to solve, but it can be intractable for thermodynamic integration. Actually, these problems are intractable for most known methods based on power posteriors in their simple form (path connecting the prior and posterior), including steppingstone sampling or annealed importance sampling ([Neal, 2001](#)).

To illustrate the problem, consider the d -dimensional parameter vector θ with a Uniform prior in the unit cube $[-0.5, 0.5]^d$, and the likelihood

$$L(\theta) = \prod_{i=1}^d \frac{1}{v\sqrt{2\pi}} \exp \left(- \frac{\theta_i^2}{2v^2} \right) + 100 \prod_{i=1}^d \frac{1}{u\sqrt{2\pi}} \exp \left(- \frac{(\theta_i - \mu)^2}{2u^2} \right). \quad (7)$$

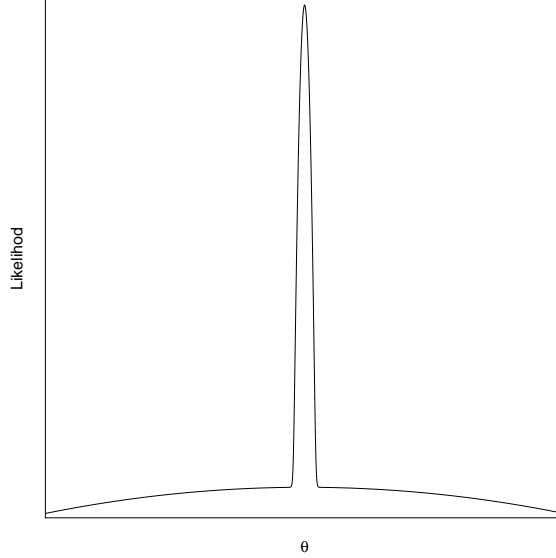


Figure 2: Likelihood function of the statistical model. This is partly convex with two phases: a plateau and a spike. The plot is not scaled but it represents didactically the likelihood function in one dimension.

Skilling (2006) considered the following values: $d = 20$, $\mu = 0$, $v = 0.1$ and $u = 0.01$. The likelihood function is the sum of two Gaussians (7): the first is a Gaussian of width 0.1 and the second Gaussian has a factor of 100 and a width 0.01 that is superposed on the first one. The likelihood function is a relatively flat density with a spike in its center. A non-scaled representation of this function is shown in Figure 2. Independently of the dimension d , the center peak μ and the variances u and v , the marginal likelihood is 101. Skilling assessed this problem analytically whereas we do it in a practical way.

We assess the marginal likelihood estimation using the harmonic mean (HM), thermodynamic integration (TI), steppingstone sampling (SS), generalized steppingstone sampling (GSS) and nested sampling (NS). To be fair in their comparison, we use around 100,000 samples for each method. For estimating TI, we use 50 transition distributions and 2000 samples from each of them. SS follows the same design, but it does not require samples from the posterior. For GSS, we consider 48 transitional distributions and 2000 posterior samples to calibrate the reference distribution. Thus, SS and GSS make the same computational effort (98,000 points in total). For nested sampling we consider 1,300 active points to make it similar in computational terms. We also include NS with only one active point.

We evaluate TI and SS under an annealing and melting scheme. Both schemes are evaluated for β values following a Beta(0.3, 1) distribution (Xie et al., 2011). For GSS, we only consider a melting scheme with β values uniformly spaced. For its reference distribution, we use the prior but reparametrized according to the posterior sample. This is done calculating the absolute maximum value θ_M (scalar) from the sample and then using it as the limits of a uniform (prior distribution). Thus, each parameter has as reference distribution a Uniform($-\theta_M, \theta_M$).

Figure 3 shows the box plots for 1,000 estimates of the log-marginal likelihood. The horizontal dotted line depicts the true value $\log(101) \approx 4.62$. HM overestimates the true value because the posterior distribution is dominated by the spike, ignoring the flat Gaussian. Under the annealing scheme, TI and SS underestimate the evidence due to the nature of the power posteriors. For some transition distributions, the power posterior consists of a mixture of the narrow and the broad gaussians where both components should contribute to

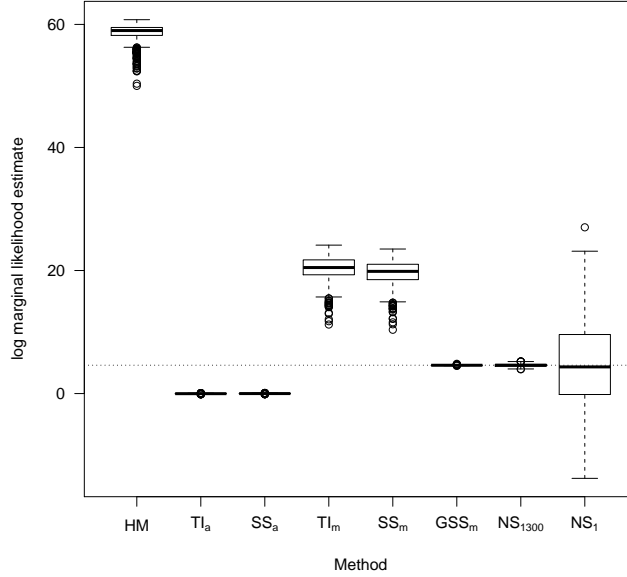


Figure 3: Performance of HM, TI, SS, GSS and NS for a partly convex likelihood function in the statistical example. The subscripts “a” and “m” depict the annealing or melting scheme used by the method, respectively. The horizontal dotted line stands for the true log marginal likelihood value.

$\mathbb{E}_{p_\beta}[\log L(\mathbf{X}|\boldsymbol{\theta}, M)]$. However, the sampler gets trapped in the wide component and cannot find the narrow component because its volume is so small. Something similar happens with TI and SS under the melting scheme, but in this case the sampler gets trapped in the spike making it unable to mix both components well. As a results, the true value is underestimated and overestimated under the annealing and melting schemes, respectively.

On the contrary, GSS and NS work perfectly. For GSS, the reference distribution encapsulates the spike and concentrate its probability mass around it. Thus the probability of going from the spike to the reference distribution (under the melting scheme) increases allowing the mixing of the two components of the likelihood. On the other hand, NS uses the likelihood contours regardless of its shape. This allows it to deal well with phase transitions. The method even works with only one active point yielding estimates around the true value. Even though NS has a higher variability, just one single run is needed to estimate also its error, unlike GSS.

4.2 Three primate example

We analyze a small data set, assuming the GTR+ Γ_4 model, to evaluate NS performance. The data comprise 15,727 sites from the mitochondrial DNA of 3 primates: human, chimpanzee and macaque. This is a subset of a data which was previously analysed in the literature (Roos et al., 2011). Even though there is not an analytical estimate of the total marginal likelihood (variable tree topology) for this simple example, it can be estimated by using all the individual marginal likelihood estimates per topology. In general, this is not viable due to the number of trees increases drastically as a function of the number of taxa. But for 3 species, there are only 3 possible rooted trees. This allows us to evaluate and compare the direct total marginal likelihood estimate. The 3 phylogenies of this example are displayed in Figure 4.

The estimated log marginal likelihood from the generalized steppingstone sampling of each phylogeny is taken as the true value to evaluate the variable tree topology analysis via nested sampling. The 3 marginal likelihoods

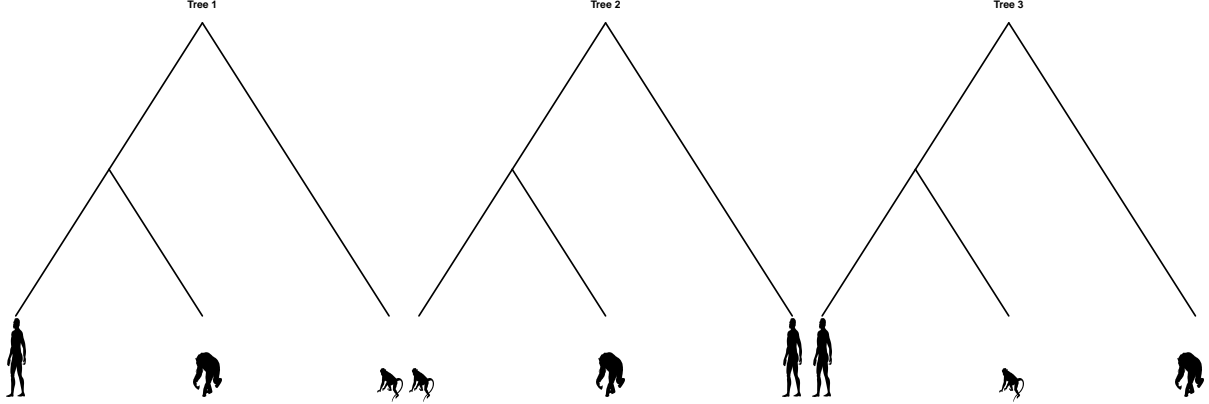


Figure 4: Three possible rooted trees for 3 primates species. According to Tree 1, the species are: human, chimpanzee and macaque.

$\log \mathcal{Z}_{\tau_1}$	$\log \mathcal{Z}_{\tau_2}$	$\log \mathcal{Z}_{\tau_3}$	$\log \mathcal{Z}$
-35669.50	-35672.36	-35672.38	-35670.49

Table 1: Estimated log marginal likelihoods per phylogeny by using generalized steppingstone sampling. These are considered as true values in order to calculate the log total marginal likelihood $\log \mathcal{Z}$.

are estimated using 500 samples from each of 200 transitional distributions. Then, the total marginal likelihood is calculated as follows

$$\mathcal{Z} = \sum_{i=1}^3 \pi(\tau_i|M) \mathcal{Z}_{\tau_i},$$

where $\mathcal{Z}_{\tau_i} = \hat{\mathcal{Z}}_{\tau_i}$ is the marginal likelihood for its respective tree τ_i , with prior probability $\pi(\tau_i|M) = 1/3$ for the given substitution model M , where $i = 1, 2, 3$. The results are presented in Table 1.

The log-total marginal likelihood has been directly estimated via NS using 50 active points. The estimate is -35670.92 with estimated standard deviation 0.88. This estimate is consistent with the value estimated by using the 3 individual tree marginal likelihoods by using GSS (see Table 1). Its uncertainty could be decreased by increasing the number of active points.

4.3 Green plant rbcL example

We study a data set previously used by Xie et al. (2011) to compare HM, TI, and SS. In Xie et al. the topology remained fixed, but we will allow for a variable topology. This data set contains 10 species of green plant. The DNA sequences are from the chloroplast-encoded large subunit of the RuBisCO gene (*rbcL*). We use this data set to assess nested sampling performance in model selection, parameter inference, and to assess its sensitivity under different prior specifications.

4.3.1 Model selection

We compare 6 models of DNA evolution: JC69, JC69+ Γ_4 , HKY85, HKY85+ Γ_4 , GTR and GTR+ Γ_4 . We consider a broad Gamma(1, 1000) distribution as prior for the gamma parameter where applicable. The total marginal likelihood estimation is carried out for each model via NS with 50 active points.

Model	$\log \mathcal{Z}$	SD
JC69	-7273.22	0.92
JC69+ Γ_4	-6917.18	1.00
HKY85	-7057.45	1.04
HKY85+ Γ_4	-6632.30	1.10
GTR	-6996.82	1.11
GTR+ Γ_4	-6617.81	1.17

Table 2: Estimated log-marginal likelihood under different substitution models for the green plant rbcL data.

Tree	1	2	3	4	5
NS	0.045	0.942	0.001	0.007	0.004
MCMC	0.042	0.945	0.000	0.009	0.004

Table 3: Proportion comparison of a tree posterior sample obtained via NS and MCMC method. Both methods yield similar distributions.

Table 2 shows the estimates and their respective uncertainties for the 6 models. The uncertainties of the estimates are relatively small with respect to the marginal likelihood differences between the models. Hence, there are no potentially overlapping confidence intervals (for instance, considering the loose interval $\log \mathcal{Z} \pm 3 \times \text{SD}$). Thus, for any model comparison, the decision in favour of the model with higher marginal likelihood would be pretty safe. The GTR+ Γ_4 has the highest log-evidence value and consequently fits the data better than the other models. Interestingly, the HKY85+ Γ_4 has evidence significantly higher than the GTR model which has 4 more rate parameters but does not include a rate across sites. This shows the importance of including a parameter that models the variability across sites for this data set.

4.3.2 Parameter inference

Model selection identified GTR+ Γ_4 as the evolutionary model best suited for these data, we make use of NS algorithm to carry out parameter inference. In particular, we focus on the tree topology. The posterior sample is obtained from the discarded points which are weighted by their contribution in the estimated evidence as described previously. We also compare this sample to another one which was obtained independently via MCMC.

Furthermore, we assess the NS tree posterior sample by calculating its effective sample size (ESS). This is a measure of the number of uncorrelated points which are needed to obtain the same information about a parameter as the one obtained from the MCMC. It is a very useful tool to assess the adequacy of posterior samples taken via MCMC analysis. However, its calculation is not direct for the tree sample. [Lanfear et al. \(2016\)](#) proposed to estimate it by randomly selecting a focal tree (from the posterior sample) and calculating the path differences of the tree sample with respect to it. Then, the ESS is calculated from these distances. Replicating this procedure provides the means to calculate a confidence interval. We use 1,000 replications to generate a 95% confidence interval.

Recycling the previous NS analysis, used previously to estimate the marginal likelihood, we infer the posterior tree distribution. The maximum number of representative samples from this single run is 814 points. The number of points required by NS in this run was of around 4,500. The sampled trees are shown in Table 3 with their respective proportions. This table also includes 1,000 posterior samples obtained independently via MCMC. The proportions are pretty similar. Both methods are in agreement inferring the posterior tree distribution. The maximum posterior tree is the same used by [Xie et al. \(2011\)](#) which was obtained by maximum likelihood method under the same evolutionary model.

Method	Prior model		
	Vague	Good	Wrong
HM	-6559.51	(-1.03)	(-35.94)
GSS	(-8.55)	-6609.04	(-59.37)
NS	(-8.42)	-6609.39	(-58.23)

Table 4: Estimated log-total marginal likelihood values for the GTR+ Γ_4 model under 3 different priors for the gamma parameter. The highest one for each method is displayed whereas the difference with the other models are shown in parentheses.

A 95% confidence interval for the ESS of this 814 tree samples is [652 – 725]. According to the criterion $ESS > 200$ (Drummond et al., 2006; Lanfear et al., 2016), the sample contains enough uncorrelated points to estimate the posterior tree distribution.

4.3.3 Sensitivity

Following the analysis carried out by Xie et al. (2011) to show that TI and SS are sensitive to prior distributions in the fixed tree topology case, unlike HM, we estimate the total marginal likelihood for the selected GTR+ Γ_4 model under three different priors. The models are just differentiated by the prior distributions placed on the shape parameter of the discrete gamma distribution of rates across sites. The remaining priors are as before. The 3 priors are the following: Gamma(1, 1000) \equiv Exponential(0.001) is the “vague” prior which has variance 1 million; Gamma(10, 0.026) is the “good” prior which is centred around the posterior mean; and Gamma(148, 0.00676) is the “wrong” prior which is centred arbitrarily at 1. The names “good” and “wrong” are just labels which are related to the relationship between the information contained in this specific data set and the priors evaluated. The “wrong” prior is in contradiction with the likelihood function (i.e. the prior density and the likelihood function peak in different regions), unlike the “good” prior. While the “vague” and the “wrong” priors seem qualitatively different, whether a prior appears vague or wrong can depend on the coordinate system. In any case, the “good” prior should lead to a better marginal likelihood.

For each of the three priors, we carry out marginal likelihood estimation allowing variable tree topology via HM, NS and GSS. For estimating HM, we use a single-chain MCMC sampler of 50,000 posterior samples. For NS, we consider 50 active points. For GSS, we use 100 transitional distributions with 500 samples from each of them. The path connecting the distributions follows an annealing scheme. NS posterior samples are used to calibrate their reference distributions (Fan et al., 2011).

Table 4 shows the estimated log-total marginal likelihood for each method under the 3 different prior distributions. As expected, HM does not discriminate between the “vague” and the “good” prior. These priors are dominated by the area of highest likelihood, unlike the “wrong” prior, which prevents the exploration of that place. On the other hand, NS and GSS are sensitive to the prior specification. These methods take into account the prior information which is reflected in the estimate. They produce the highest and lowest values for the models with the good and wrong prior, respectively. Thus, NS, and also GSS, impose more penalty to those models which possess unnecessary parameters, but in a way that depends on how much is known about those extra parameters, and whether they affect the model fit.

4.4 Laurasiatherian data

Laurasiatheria is a superorder of placental mammals which originated on the northern supercontinent of Laurasia. The data set contains 47 RNA sequences of length 3,179 from this group. Each aligned sequence represents a different species. The data is contained in the R-package *phangorn* (Schliep, 2011).

This data set presents two interesting and challenging characteristics which provide a good scenario to assess

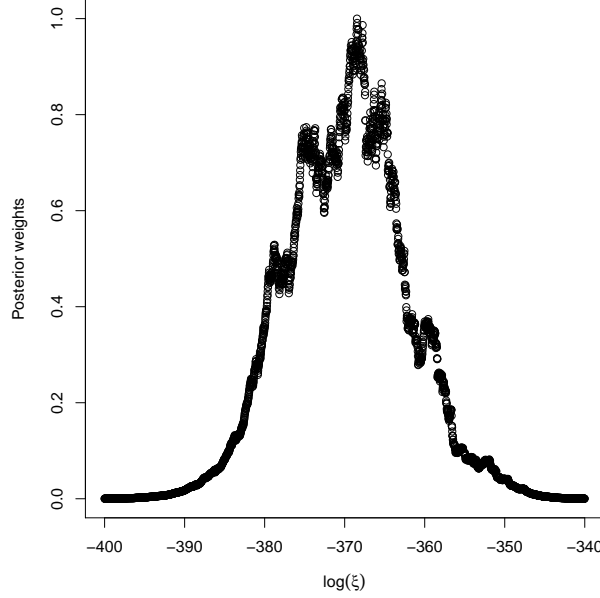


Figure 5: Posterior weights for the Laurasiatherian data. The NS posterior sample is taken according to these weights.

nested sampling performance. Firstly, the tree space contains around $3.529995e+68$ possible phylogenies that NS has to potentially explore. Secondly, the posterior distribution is not dominated by a single tree topology. This represents a challenge for MCMC methods.

Assuming a GTR+ Γ_4 model and a Gamma(2, 2.5) distribution as a prior for the gamma parameter, we use NS in order to sample the posterior distribution. Actually, we are only interested in studying the tree posterior distribution. We assess this sample according to 2 analyses proposed by [Lanfear et al. \(2016\)](#). For the first analysis, we calculate the path difference between the tree samples and a focal tree. We consider the tree with the highest frequency as the focal tree. Then, these differences are visualized in a trace plot. For the second analysis, we estimate a confidence interval for the effective sample size (ESS) as described in the previous example.

NS, with 50 active points, yielded in this single run a sample of 1,525 points. It required around 20,000 points to explore the parameter space to estimate the marginal likelihood. The posterior weights of the discarded points are shown in Figure 5. This gives a quick picture of the posterior density. The posterior is concentrated in a small area of the prior (in around e^{-340} of its mass). This is mainly caused by the uniform prior assigned to the huge tree space. The trace plot based on the path differences is shown in Figure 6. The trace is pretty stable seeming that the chain has reached the equilibrium distribution, i.e. the posterior. There are no signs of autocorrelation and it appears to have a good mixing. Note that two or more different tree topologies can have the same path difference respect to the focal tree making the trace seem at times constant despite a change in state. The 95% confidence interval for the ESS is [1494 – 1525]. A rule of thumb suggests 200 to be a lower limit to accurately infer the tree posterior distribution ([Drummond et al., 2006](#); [Lanfear et al., 2016](#)).

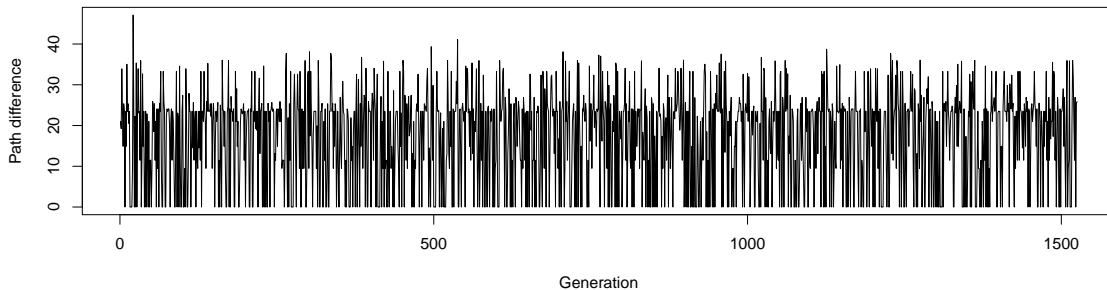


Figure 6: Path differences in nested sampling tree posterior sample respect to the maximum posterior tree for the Laurasiatherian data. The trace shows reasonably good mixing.

5 Conclusion

Nested sampling is a general Bayesian algorithm that provides the means to estimate marginal likelihoods and to carry out parameter inference. We have introduced it to phylogenetic, allowing variable tree topology. Its performance has been compared to established methods available in many phylogenetic software packages.

Many of the methods currently used in phylogenetics do not work under certain scenarios. The problems related to HM are well documented, unlike those related to power posterior methods. These approaches are often presented as methods of general applicability but they also possess limitations. For instance, in the statistical example presented above, TI and SS would require impractical computational effort to yield an acceptable estimate. This is an extreme case, but phase changes can also be found in some data analysis problems (e.g. [Brewer and Donovan, 2015](#)). On the other hand, GSS and NS are able to deal well with these likelihood shapes. With a reasonable reference distribution, GSS works well, but NS even in poor conditions, i.e. one active point ($N = 1$), yields estimates around the true value.

We have assessed NS performance in phylogenetics. Firstly, we tested it in a small data set of 3 primates. This allowed us to calculate the total marginal likelihood indirectly by averaging all the individual marginal likelihood estimates per tree topology. In this manageable scenario, NS is consistent with this estimate. Secondly, we have carried out model selection among 6 DNA substitution models for a data set which contains 10 species of green plant. After selecting the evolutionary model, we have compared the posterior tree distribution yielded by NS with one which was obtained independently via MCMC. The similarity is clear. Also, we have shown that NS, like GSS, is sensitive to prior specifications. Finally, we have made use of NS to analyze the Laurasiatherian group and generate samples from its bimodal posterior distribution. The method yields more than an acceptable effective sample size using 50 active points.

TI and SS have become popular because of their high accuracy estimating the marginal likelihood. Specifically, SS has become popular due to its implementation in widely used phylogenetic software. However, they rely on several problem-specific tuning parameters which need to be specified by the user. Namely, number and distribution of the β values and a number of samples from each transitional distribution. GSS dispenses with the distribution of the β values, but it requires a number of posterior samples to calibrate the reference distribution. This calibration is essential to get good estimates. On the other hand, NS only requires the number of active points and the number of MCMC steps used to generate a replacement point. The length of the run is determined by the termination conditions described previously. NS is in practice more user-friendly than the methods presented.

NS not only yields a marginal likelihood estimate but also provides a measure of its uncertainty, which is inversely proportional to the square root of the number of active points. The higher this number, the more accurate the estimate. Under similar computational conditions, NS can have a higher uncertainty than GSS,

though this depends on the problem and whether the GSS parameters are well tuned. However, its uncertainty can be calculated directly in a single run whereas GSS requires several replications to estimate its uncertainty. In practice, one could estimate a marginal likelihood interval for the competitive models via NS, and then check if they overlap. If so, the method could be run again for those models with more active points to guarantee a higher precision.

Nested sampling also provides the means to carry out parameter inference. This does not involve an extra cost since the samples used to estimate the marginal likelihood are recycled. In particular, we have assessed the method to study posterior tree distributions. In the data analysed, only 50 active points were more than enough to infer the posterior according to the effective sample size criteria. The posterior sample size is directly proportional to the number of active points. Unlike conventional MCMC methods, NS does not require a burn-in period. In general, this period represents a high computational cost for MCMC methods.

NS provides several positive characteristics which make it competitive to established methods in phylogenetics. It has been applied successfully to different fields and we believe this success can be replicated in phylogenetics.

References

- Aitken, S. and O. Akman. 2013. Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. *BMC Syst. Biol.* 7:72.
- Arima, S. and L. Tardella. 2014. Inflated density ratio (IDR) method for estimating marginal likelihoods in Bayesian phylogenetics. chap. 3, Pages 25–58 *in* Bayesian phylogenetics : methods, computational algorithms, and applications (M. Chen, L. Kuo, and P. O. Lewis, eds.). Chapman and Hall/CRC, New York.
- Baele, G. and P. Lemey. 2014. Bayesian model selection in phylogenetics and genealogy-based population genetics. chap. 4, Pages 59–94 *in* Bayesian phylogenetics : methods, computational algorithms, and applications (M. Chen, L. Kuo, and P. O. Lewis, eds.). Chapman and Hall/CRC, New York.
- Baele, G., P. Lemey, and S. Vansteelandt. 2013. Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. *BMC Bioinformatics* 14:85.
- Brewer, B. J. and C. P. Donovan. 2015. Fast Bayesian inference for exoplanet discovery in radial velocity data. *Monthly Notices of the Royal Astronomical Society* 448:3206–3214.
- Brewer, B. J., L. B. Pártyay, and G. Csányi. 2011. Diffusive nested sampling. *Stat. Comput.* 21:649–656.
- Chopin, N. and C. Robert. 2010. Properties of nested sampling. *Biometrika* 97:741–755.
- Drummond, A. J. and R. Bouckaert. 2015. Bayesian evolutionary analysis with BEAST. Cambridge University Press.
- Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
- Drummond, A. J. and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214.
- Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29:1969–1973.
- Fan, Y., R. Wu, M.-H. Chen, L. Kuo, and P. O. Lewis. 2011. Choosing among partition models in Bayesian phylogenetics. *Mol. Biol. Evol.* 28:523–532.
- Feroz, E., M. P. Hobson, and M. Bridges. 2009. Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Mon. Not. R. Astron. Soc.* 398:1601–1614.

- Friel, N. and A. N. Pettitt. 2008. Marginal likelihood estimation via power posteriors. *J. Roy. Stat. Soc. B* 70:589–607.
- Handley, W. J., M. P. Hobson, and A. N. Lasenby. 2015. POLYCHORD: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society* 453:4384–4398.
- Holder, M., P. O. Lewis, D. L. Swofford, and D. Bryant. 2014. Variable tree topology stepping-stone marginal likelihood estimation. chap. 5, Pages 95–111 *in* Bayesian phylogenetics : methods, computational algorithms, and applications (M. Chen, L. Kuo, and P. O. Lewis, eds.). Chapman and Hall/CRC, New York.
- Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. Evol.* 21:1123–1133.
- Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
- Kass, R. E. and A. E. Raftery. 1995. Bayes factors. *J. Amer. Statist. Assoc.* 90:773–795.
- Knuth, K. H. and J. Skilling. 2012. Foundations of inference. *Axioms* 1(1):38–73.
- Lanfear, R., X. Hua, and D. L. Warren. 2016. Estimating the Effective Sample Size of tree topologies from Bayesian phylogenetic analyses. *Genome Biol. Evol.* 8:2319–2332.
- Larget, B. and D. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16:750–759.
- Lartillot, N., T. Lepage, and S. Blanquart. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.
- Lartillot, N. and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration. *Syst. Biol.* 55:195–207.
- Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* 24:2669–2680.
- MacKay, D. J. C. 2002. Information theory, inference & learning algorithms. Cambridge University Press, New York, NY, USA.
- Mukherjee, P., D. Parkinson, and A. R. Liddle. 2006. A nested sampling algorithm for cosmological model selection. *Astrophys. J. Lett.* 638:L51–L54.
- Neal, R. M. 2001. Annealed importance sampling. *Stat. Comput.* 11:125–139.
- Newton, M. A. and A. E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B* 56:3–48.
- Pullen, N. and R. J. Morris. 2014. Bayesian model comparison and parameter inference in systems biology using Nested Sampling. *PLoS ONE* 9:e88419.
- R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.
- Rannala, B. and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* 43:304–311.
- Rannala, B., T. Zhu, and Z. Yang. 2011. Tail paradox, partial identifiability and influential priors in Bayesian branch length inference. *Mol. Biol. Evol.* 29:325–335.

- Roos, C., D. Zinner, L. S. Kubatko, C. Schwarz, M. Yang, D. Meyer, S. D. Nash, J. Xing, M. A. Batzer, M. Brameier, F. H. Leendertz, T. Ziegler, D. Perwitasari-Farajallah, T. Nadler, L. Walter, and M. Osterholz. 2011. Nuclear versus mitochondrial DNA: evidence for hybridization in colobine monkeys. *BMC Evol. Biol.* 11:77.
- Schliep, K. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593.
- Sivia, D. S. and J. Skilling. 2006. *Data analysis: a Bayesian tutorial*. Oxford University Press, USA.
- Skilling, J. 2006. Nested sampling for general Bayesian computation. *Bayesian Analysis* 1:833–860.
- Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time markov chain evolutionary models. *Mol. Biol. Evol.* 18:1001–1013.
- Verdinelli, I. and L. Wasserman. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Amer. Statist. Assoc.* 90:614–618.
- Walter, C. 2014. Point Process-based Monte Carlo estimation. *ArXiv e-prints* .
- Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* 60:150–160.
- Yang, Z. and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.* 14:717–724.